Setting Pass Scores for Clinical Skills Assessment

Min Liu and Keh-Min Liu
Department of Anatomy, College of Medicine, Kaohsiung Medical University, Kaohsiung, Taiwan.

In a clinical skills assessment, the decision to pass or fail an examinee should be based on the test content or on the examinees' performance. The process of deciding a pass score is known as setting a standard of the examination. This requires a properly selected panel of expert judges and a suitable standard setting method, which best fits the purpose of the examination. Six standard setting methods that are often used in clinical skills assessment are described to provide an overview of the standard setting process.

Key Words: clinical skills assessment, standard setting
(*Kaohsiung J Med Sci* 2008;24:656–63)

According to Harden, the five main functions of clinical assessments are: (1) to pass or fail the examinee; (2) to grade the examinee; (3) to provide feedback to the examinee; (4) to provide feedback to the examiner; and (5) to motivate the examinee [1]. Because there are always examinees whose scores are not in the high-score pass group or in the low-score fail group, examiners must decide "How much is enough?" to pass an examination [2]. This process of deciding a pass score is known as setting the standard of the examination.

While many examiners choose 60% as the pass score, this decision is usually based on tradition rather than on test content or on examinees' performance. This may impose the possibility of either failing examinees who have achieved the required level of knowledge and skills if the standard is too high, or passing examinees who have not achieved the requirements if the standard is too low [2]. Examiners may also find it difficult to provide a defensible explanation of how this 60% passing standard was set.

Kane et al emphasized that there is no "gold or correct standard" to be discovered [3,4]. Instead, the pass score is a consensus of a panel of expert judges (or standard setters) regarding the cut-point, which is determined using a systematic judgmental method [5]. Therefore, selecting different expert judges or using different standard setting methods, will result in different pass scores for the same examination [6].

Relative Standards *vs.* Absolute Standards

Standards fall into two types, relative (norm-referenced) and absolute (criterion-referenced), based on the object for comparison. If the standard is based on the performance compared between the score distribution of the whole examinee group and an examinee's score, it is a relative standard [2,6,7]. For example, an examinee whose score is in the upper 60% of the group will pass, or an examinee whose score is less than 1.5 standard deviations below the mean score will fail. Whether an examinee passes or fails the test depends on how much he/she scores relative to the

Received: Jan 5, 2009 Accepted: Jan 22, 2009
Address correspondence and reprint requests to: Dr Keh-Min Liu, College of Medicine, Kaohsiung Medical University, 100 Shin-Chuan 1st Road, Kaohsiung 807, Taiwan.
E-mail: kemili@kmu.edu.tw

scores of his/her peer examinees. Relative standards are often used in low-stakes examinations or examinations to select a limited number of examinees, such as for admissions [5,8].

If the standard is based on the content of the test and is independent of the performance of the whole examinee group, it is an absolute standard [2,6,7]. For example, an examinee who answers 75% of the questions correctly will pass. Whether an examinee passes or fails the test depends on his/her level of knowledge and skills compared with clearly-defined criteria [9]. Absolute standards are often used in high-stakes examinations such as final examinations or graduating examinations, where "how much does an examinee actually know or can do" is critical [5,8].

The decision to use a relative or an absolute standard should be based on the purpose of the examination [5]. However, Case and Swanson suggested that "Unless there are strong reasons to fail a given number of examinees, an absolute standard (based on examinee performance) is preferred over a relative standard (based on a particular failure rate)" [7]. Turnbull also stated that "As the principal objective of medical education is to produce a competent physician, then, unquestionably, the *raison d'être* for the evaluation process is to assess the competence and not the rank order of students" [9]. Therefore, an absolute standard that is based on carefully established criteria will be more appropriate to determine whether or not an examinee has achieved the required level of competence [10].

Absolute standards can be further classified into test-centered methods and examinee-centered methods based on the subject of judgment.

Test-centered Methods

Test-centered methods are based on judgments about test questions or, in an objective structured clinical examination (OSCE), the checklist items. The judges review the items and determine the expected performance, which indicates that a "borderline" examinee has obtained the minimal competence required to pass the test. A borderline examinee is an examinee whose performance is on the borderline between the upper group (those whose performance are good enough to pass) and the lower group (those whose performance are not good enough and fail), and he/she has an equal chance to barely pass or fail the test. Therefore, the pass score should be the score expected of a borderline examinee [2,6,11]. A test-centered standard-setting process comprises five steps [2,6,12]:

1. Judge selection: Select judges who are experts in the content area of the test, are familiar with the levels of knowledge and skills of the examinees, and understand how borderline examinees will perform on test items.
2. Orientation: Provide the judges with information such as the content, format and purpose of the OSCE, case materials and checklists of each station, and the standard-setting method.
3. Define borderline examinees: The judges discuss and agree upon the characteristics of borderline examinees and their knowledge and skill levels with specific examples. The borderline examinees defined in this step are based not on actual performance, but on what the judges perceive as borderline performance; therefore, they are hypothetical borderline examinees.
4. Make judgments: The judges use one of the test-centered standard-setting methods such as the Angoff method or the Ebel method to make judgments. For each item, each judge makes an independent judgment of "how will a hypothetical borderline examinee perform on that item?" with the difficulty and the importance of the item in mind. They can discuss and change their judgments during this step.
5. Set the pass score: The score of an item is determined by averaging the scores of that item given by each judge. The pass score is then calculated by summing up the scores for all items.

Two test-centered methods, the Angoff method and the Ebel method, are described below. The first three steps are the same for both methods.

Angoff method

The Angoff method requires the judges to make judgments on "How well will a hypothetical borderline examinee perform on each item?" The judges decide for each item "the probability that a borderline examinee will perform that item correctly" or, alternatively, "the percentage of borderline examinees who will perform that item correctly" [2,5,6,12]. The pass score is then calculated by summing up the probabilities or percentages for all items (Column 3 in Table 1).

Table 1. Example of setting the pass score of a 10-item station using the Angoff method

Item number	% of all examinees who performed the item correctly	Angoff method*	Yes/No Angoff method†	Three-level (Yes/No/Maybe) Angoff method†
1	100%	0.90	1 (Yes)	1 (Yes)
2	85%	0.75	1 (Yes)	1 (Yes)
3	65%	0.50	0 (No)	0 (No)
4	75%	0.60	1 (Yes)	0.5 (Maybe)
5	95%	0.85	1 (Yes)	1 (Yes)
6	50%	0.45	0 (No)	0 (No)
7	75%	0.60	1 (Yes)	0.5 (Maybe)
8	70%	0.55	0 (No)	0.5 (Maybe)
9	90%	0.75	1 (Yes)	1 (Yes)
10	80%	0.75	1 (Yes)	1 (Yes)
Pass score	–	6.7	7	6.5

*Probability that a borderline examinee will perform that item correctly (pass score = sum of probabilities for all items); †Yes = 1, Maybe = 0.5, No = 0 (pass score = sum of all item codes). Modified from Yudkowsky et al [13].

Because it may not be easy for the judges to make decisions in terms of probability, two simplified Angoff methods can be used: the Yes/No Angoff method and the three-level Yes/No/Maybe Angoff method [6,13]. For each item, judges are asked the following question "Will a borderline examinee perform that item correctly?" In the Yes/No Angoff method, the answer "Yes, he/she will" is coded as 1, while the answer "No, he/she won't" is coded as 0 (Column 4 in Table 1). In the three-level Angoff method, an additional answer "Maybe, he/she has a 50:50 chance" is coded as 0.5 (Column 5 in Table 1). The sum of all item codes is the pass score.

Although the task is to estimate the real performance of examinees (i.e. how well examinees will perform) and not the expected performance of examinees (i.e. how well examinees should perform) [6], judges tend to set the pass scores unrealistically high. Providing them with actual performance data such as the mean score and standard deviation or "the percentage of all examinees who performed the item correctly" will help them understand the difficulty of the items and moderate their judgments [6,12] (Column 2 in Table 1).

Ebel method

The Ebel method includes two rounds of judgments, which can be organized into a classification table like that in Table 2 [2,6,11,14]. In the first round, the judges classify all of the checklist items into 12 categories based on three levels of difficulty (easy, medium, hard) and four levels of relevance (essential, important, acceptable, questionable). The definition of each level of difficulty and relevance should be discussed in advance. Actual performance data such as "the percentage of all examinees that performed the item correctly" will assist the judges in deciding the level of difficulty for an item. In the second round, the judges estimate for each category "the percentage of items that a borderline examinee will perform correctly". Multiplying the number of items in a category by the percentage correct gets the score for that category. The pass score is then calculated by summing up the scores for the 12 categories. This method is more complicated than other test-centered methods, and therefore involves more time and effort from experienced judges [6].

Examinee-centered Methods

Examinee-centered methods are based on judgments about individual examinees, and not on the test items or the test scores. The judges need to observe the actual performance of examinees and categorize them into groups according to their level of mastery. The pass score is set at the score, which best describes the judgments, i.e. examinees whose scores are above the pass score have actually performed well enough to pass, while examinees whose scores are below the pass score really have not performed well enough and will fail [3,8,12]. Although reviewing the performance of every examinee is time consuming, judges may find examinee-centered methods easier to perform. This is

Table 2. Example of setting the pass score of a 10-item station using the Ebel method

Relevance	Difficulty								
	Easy			Medium			Hard		
	Item number	(number of items in category)	% of items correct	Item number	(number of items in category)	% of items correct	Item number	(number of items in category)	% of items correct
Essential (score for category*)	1,5	(2) (2×0.95=1.90)	95%	10	(1) (1×0.85=0.85)	85%	7	(1) (1×0.60=0.60)	60%
Important (score for category*)	9	(1) (1×0.90=0.90)	90%	4	(1) (1×0.75=0.75)	75%	3	(1) (1×0.55=0.55)	55%
Acceptable (score for category*)	2	(1) (1×0.85=0.85)	85%	8	(1) (1×0.65=0.65)	65%	N/A		N/A
Questionable (score for category*)	N/A		N/A	N/A		N/A	6	(1) (1×0.30=0.30)	30%

*Score for category = (number of items in category)×(% of items correct). Pass score = sum of scores for 12 categories = 1.90+0.85+0.60+0.90+0.75+0.55+0.85+0.65+0.30 = 7.35.
Modified from Downing et al [6].

because they are more familiar with making direct judgments on actual performance of real examinees, rather than estimating "the probable performance of a hypothetical group (of examinees)", as stated by Cizek [5,8,15]. Three examinee-centered methods—the contrasting groups method, the borderline group method and the borderline regression method—are described below.

Contrasting groups method

The contrasting groups method requires the judges to observe the performance of examinees and divide them into two groups: pass (qualified) and fail (unqualified). The pass score is set at the score, which maximizes the discrimination between qualified and unqualified examinees [2,6,12]. This standard-setting process comprises five steps [2,6,8]:

1. Judge selection: The judges must be experts who know the level of performance required to pass the test, and are able to determine each examinee's performance level.
2. Orientation: Provide the judges with information about the OSCE, materials of the station under review, the standard-setting method and, most importantly, a clear definition of "qualified" examinees.
3. Training: The judges are trained to observe and rate the performance that the station is intended to assess. Video samples of different levels of mastery can help judges understand the range of performances they might encounter.
4. Observe and make judgments: The judges observe each examinee and give a global pass/fail rating of the examinee's overall performance. The judges should remember that their judgments should be based on the actual performance of an examinee, not the examinee's checklist score.
5. Set the pass score: The examinees are divided into a pass group and a fail group according to the global rating results. The checklist score distributions of the two groups are drawn, and the point where the two distributions intersect is set as the pass score (Figure 1).

Borderline group method

The borderline group method is a similar method, in which the judges identify a third group of examinees, the borderline group, whose performance just meets the required level of mastery [3]. The checklist scores

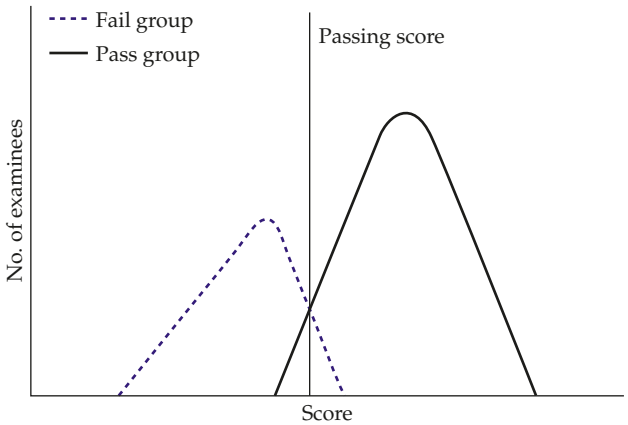

Figure 1. *Contrasting Groups Method (modified from Downing et al [6]).*

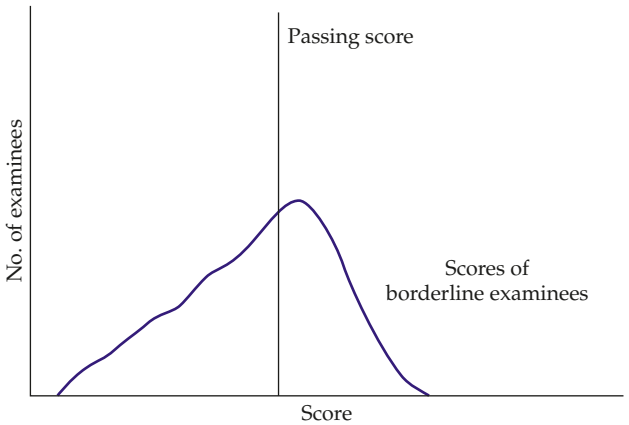

Figure 2. *Borderline Group Method (modified from Downing et al [6]).*

of the borderline group are used to set the pass score. This standard-setting process is also comprised of five steps [2,6,8]:

1. Judge selection: as above.
2. Orientation: Provide the judges with the same information listed above and the definition of "borderline" examinee. Boulet et al recommended that judges define the borderline group by identifying "those performances where they (the judges) are unsure as to whether the examinee is qualified or unqualified" [8].
3. Training: as above.
4. Observe and make judgments: The judges observe each examinee and give a global pass-borderline-fail rating of the overall performance of the examinee.
5. Set the pass score: The borderline examinees are identified and their checklist scores collected. The mean score for this group is set as the pass score (Figure 2).

Borderline regression method

If the number of borderline examinees is small, using the borderline group method will set a pass score with a higher statistical error [16]. The borderline regression method uses the checklist scores and the global rating scores of all examinees to set a pass score and is more suitable, and easier to perform in this situation. This standard-setting process comprises five steps [16–20]:

1. Judge selection: as above.
2. Orientation: Provide the judges with the same information listed above and clear definition of each point in the global rating scale (i.e. fail = 1, borderline = 2, pass = 3, very good = 4).
3. Training: as above.
4. Observe and make judgments: The judges observe each examinee and give a global rating of the overall performance of the examinee. There are two methods in this step: (1) two judges observe an examinee, one checks the checklist while the other gives a global rating; (2) one judge checks the checklist and gives a global rating. The second method is easier to apply, but the checklist score may influence the global rating score. Clear instruction and sufficient training on global rating is essential for this method.
5. Set the pass score: Checklist scores and the global rating scores of all examinees for the station are used to produce a regression equation ($y = a + bx$). The scale representing the borderline group ($x = 2$) is then inserted into the equation to calculate the checklist pass score [19] (Figure 3).

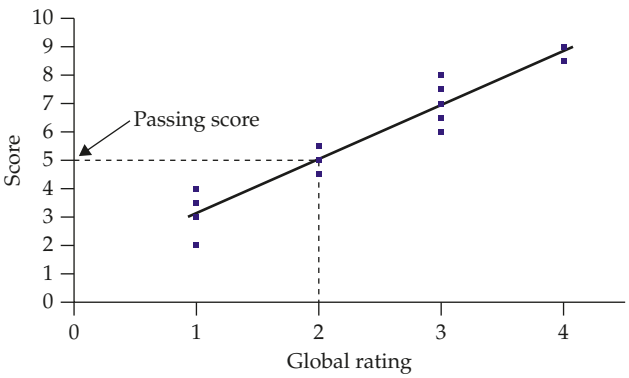

Figure 3. *Borderline Regression Method global rating: 1 = fail; 2 = borderline; 3 = pass; 4 = very good (modified from Onishi [19]).*

Relative-absolute Compromise Methods

The Hofstee method is one of the relative-absolute compromise methods, which has characteristics of both types of standards. Similar to setting a relative standard, the judges must decide the percentage of examinees to fail. Similar to setting an absolute standard, the judges must decide the pass score based on the level of performance expected of the examinees [5]. The Hofstee standard-setting process comprises four steps [6,7,12]:

1. Judge selection: The judges must be content experts who are familiar with the examinees' levels of knowledge and skills.
2. Orientation: Provide the judges with information on the OSCE, actual performance data, and the standard-setting method.
3. Make judgments: After reviewing the case materials, checklists, scoring methods and the actual performance data of a station, the judges need to make four decisions for the station: (1) the lowest acceptable percentage of examinees to fail (minimum failure rate); (2) the highest acceptable percentage of examinees to fail (maximum failure rate); (3) the lowest score to allow examinees to pass (minimum pass score); and (4) the highest score required for examinees to pass (maximum pass score).
4. Set the pass score: The actual performance data of the whole examinee group are used to draw a cumulative frequency curve of the test score for the station (Figure 4). The two failure rates and the two pass scores are used as four points to create a rectangle on the cumulative score graph, and a diagonal line can be drawn from upper left to lower right. The point where the curve and the diagonal line intersect is set as the pass score.

Compensatory Standard or Conjunctive Standard

After using one of the methods described above to set the pass scores for each station, the judges must decide on the method to set the overall pass score of the whole OSCE. For example, in a five-station OSCE in which the pass scores are 7, 6, 6.5, 8 and 5.5, the judges may decide to use the sum of these five pass scores (i.e. 33) as the overall pass score. This compensatory standard allows any examinee whose total score is higher than 33 to pass the OSCE, even if he/she failed some stations. Alternatively, the judges may decide on: (1) the number of stations, for example four out of five stations, that examinees must pass to pass the OSCE; or (2) the required stations, for example stations three and five, that examinees must pass to pass the OSCE. This conjunctive standard does not allow an examinee who failed two stations (or who failed the required stations) to pass the OSCE, even if he/she earned the highest scores in the other three stations and had a high total score [12,19,21].

Discussion

The choice of standard setting method should be consistent with the purpose of the assessment, supported by published research, easy to implement with the

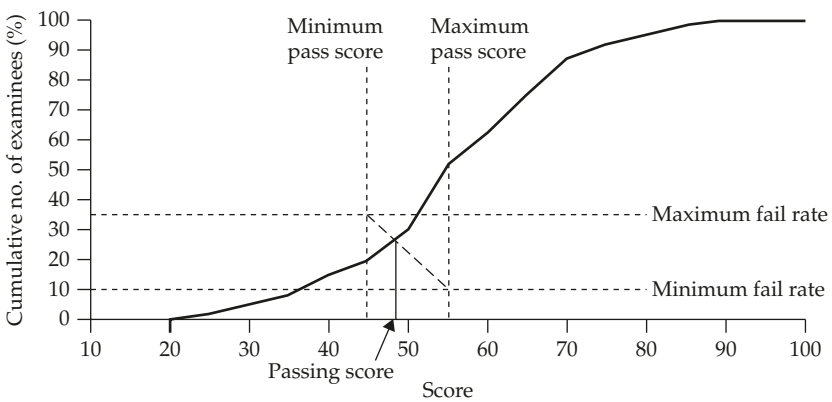

Figure 4. *Hofstee Method (modified from Downing et al [6]).*

resources available, and transparent to explain the rationale for this pass score [2,5,22]. If experts are available to actually observe examinees performance, examinee-centered methods such as the contrasting groups method or borderline group method are preferred. Categorizing the performance of examinees is a more intuitive task for most experts [2,6,8]. If it is not possible for experts to observe the performance of examinees, test-centered methods or relative-absolute compromise methods can be used. However, actual performance data should be provided for judges to make informed decisions and prevent unrealistic results [2,6,22].

Regardless of the standard setting method used, multiple experts and resources are required to perform this arduous task. Although asking each case-author to decide a pass score for the OSCE case he/she wrote in advance might make things easier, Humphrey-Murto and MacFadyen did not recommend this method [23]. Their research found that case-authors tend to set higher pass scores when compared with pass scores determined using the modified borderline group method. They implied that this may be due to case-authors, who are content experts and/or clerkship coordinators, often expecting examinees to perform better in their expert field, thereby setting higher pass scores. Involving multiple experts from different backgrounds will ensure that a variety of perspectives are considered when making the final decision [22].

Medical schools, educators and examiners have an obligation to ensure that all examinees who have passed an assessment are competent in providing the required level of health care. Therefore, selecting an appropriate panel of experts and using an appropriate standard setting method to set pass scores based on clearly stated criteria, will assist them in providing evidence that the decision of who is competent or not is based on a transparent, fair and defensible process [24].

Acknowledgments

We would like to thank Dr Win May (Keck School of Medicine of the University of Southern California, USA) and D. Hirotaka Onishi (International Research Center for Medical Education, University of Tokyo, Japan) for reviewing this article.

References

1. Harden RM. Assess clinical competence—an overview. *Med Teach* 1979;1:289–96.
2. Livingston SA, Zieky MJ. *Passing Scores: A Manual for Setting Standards of Performance on Educational and Occupational Tests.* Princeton, NJ: Educational Testing Service, 1982.
3. Kane M. Choosing between examinee-centered and test-centered standard-setting methods. *Educ Assess* 1998;5:129–45.
4. Kane MT, Crooks TJ, Cohen AS. Designing and evaluating standard-setting procedures for licensure and certification tests. *Adv Health Sci Educ* 1999;4:195–207.
5. Norcini JJ. Setting standards on educational tests. *Med Educ* 2003;37:464–9.
6. Downing SM, Tekian A, Yudkowsky R. Procedures for establishing defensible absolute passing scores on performance examinations in health professions education. *Teach Learn Med* 2006;18:50–7.
7. Case SM, Swanson DB. *Constructing Written Test Questions for the Basic and Clinical Sciences.* Philadelphia: National Board of Medical Examiners, 2001.
8. Boulet JR, De Champlain AF, McKinley DW. Setting defensible performance standards on OSCEs and standardized patient examinations. *Med Teach* 2003;25:245–9.
9. Turnbull JM. What is … Normative versus criterion-referenced assessment. *Med Teach* 1989;11:145–50.
10. McAleer S. Formative and summative assessment. In: Dent JA, Harden RM, eds. *A Practical Guide for Medical Teachers*, 1st edition. Edinburgh: Churchill Livingstone, 2001:293–302.
11. Downing SM, Lieska NG, Raible MD. Establishing passing standards for classroom achievement tests in medical education: a comparative study of four methods. *Acad Med* 2003;78:S85–7.
12. Friedman Ben-David M. AMEE Guide No. 18: Standard setting in student assessment. *Med Teach* 2000;22:120–30.
13. Yudkowsky R, Downing SM, Popescu M. Setting standards for performance tests: a pilot study of a three-level Angoff method. *Acad Med* 2008;83:S13–6.
14. Cusimano MD, Rothman AI. The effect of incorporating normative data into a criterion-referenced standard setting in medical education. *Acad Med* 2003;78:S88–90.
15. Cizek GJ. An NCME instructional module on: setting passing scores. *Educ Meas: Issues Pract* 1996;15:20–31.
16. Wood TJ, Humphrey-Murto SM, Norman GR. Standard setting in a small scale OSCE: a comparison of the Modified Borderline-Group Method and the Borderline Regression Method. *Adv Health Sci Educ* 2006;11:115–22.
17. Kramer A, Muijtjens A, Jansen K, et al. Comparison of a rational and an empirical standard setting procedure for an OSCE. *Med Educ* 2003;37:132–9.
18. Hobma SO, Ram PM, Muijtjens AM, et al. Setting a standard for performance assessment of doctor-patient communication in general practice. *Med Educ* 2004;38: 1244–52.

19. Onishi H. Theories of assessment methods in OSCE. In: Otaki J, ed. *Theories and Practices of OSCE*. Tokyo: Shinohara-shuppan Shinsha, 2007:18–39. [In Japanese]
20. Boursicot KA, Roberts TE, Pell G. Using borderline methods to compare passing standards for OSCEs at graduation across three medical schools. *Med Educ* 2007;41:1024–31.
21. Hambleton RK. *Setting Standard on Performance Assessments: Promising New Methods and Technical Issues.* Paper presented at the annual meeting of the American Psychological Association, New York, August 1995. Available from http://eric.ed.gov/ [Date accessed: November 20, 2008]
22. Norcini JJ, Shea JA. The credibility and comparability of standards. *Appl Meas Educ* 1997;10:39–59.
23. Humphrey-Murto S, MacFadyen JC. Standard setting: a comparison of case-author and modified borderline-group methods in a small-scale OSCE. *Acad Med* 2002;77:729–32.
24. Searle J. Defining competency—the role of standard setting. *Med Educ* 2000;34:363–6.